Cell Stem Cell
Protocol Review

Browsing (Epi)genomes: A Guide to Data Resources and Epigenome Browsers for Stem Cell Researchers

Rahul Karnik[1,2,3] and Alexander Meissner[1,2,3,*]
[1]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[3]Harvard Stem Cell Institute, Cambridge, MA 02138, USA
*Correspondence: alexander_meissner@harvard.edu
http://dx.doi.org/10.1016/j.stem.2013.06.006

Over the past years we have witnessed an explosion in the generation of freely available genome-wide data sets, including maps of various histone modifications, transcription factor binding, DNase hypersensitivity, and DNA methylation, which provide valuable resources for data validation, exploration, and hypothesis generation. The goal of this review is to provide the reader with information on where to find many of the data sets and how to utilize the various (epi)genome browsers for display and initial analysis. We provide selected examples to highlight key features and demonstrate the application of these browsers to stem cell biology.

Introduction

Deciphering the complexity of mammalian genome regulation is a daunting task, but it has gained notable traction with the revolution in sequencing technologies over the past few years. In particular, the ability to efficiently map diverse epigenetic modifications on a genome-wide scale has helped to annotate the genome further and pinpoint regions of potential functional significance (Barski et al., 2007; Mikkelsen et al., 2007; Ernst et al., 2011; Dunham et al., 2012). The power of epigenomic data comes from dynamics observed across cell types and as a result it will continue to increase as more data are generated. Fortunately, most of these data are available through public databases (Chadwick, 2012) and with the help of genome browsers can be readily accessed even by noncomputational researchers.

Stem cells are a popular system to study mechanisms of epigenetic regulation because they can be induced to differentiate into several lineages, which results in globally orchestrated epigenetic dynamics (Mikkelsen et al., 2007; Pan et al., 2007; Meissner, 2010; Wamstad et al., 2012; Gifford et al., 2013; Xie et al., 2013). Moreover, stem cells are a convenient system to dissect the role of epigenetic regulators due to the ease of manipulation in embryonic stem cells (ESCs) and the existence of nonlethal knockout lines, in stark contrast to the essential role of these regulators in somatic cells (Meissner, 2010; Smith and Meissner, 2013). As a result, a significant number of the already available data sets cover a wide range of undifferentiated ESCs and their derivatives in mouse and human (Mikkelsen et al., 2007; Pan et al., 2007; Meissner, 2010; Wamstad et al., 2012; Gifford et al., 2013; Xie et al., 2013).

The creation of genome browsers such as UCSC and Ensembl has greatly facilitated the widespread use of the reference genomes and has certainly increased the impact of genome projects (Hubbard et al., 2002; Kent et al., 2002). These browsers have been adopted for the display of epigenome data and similarly accelerate access as well as use. However, the increasingly complex nature of epigenomic data sets (many modifications that are uniquely distributed in every cell type) has spurred the development of newer multidimensional epigenome browsers and additional integrated tools. In this review, we will briefly introduce key sources for obtaining public data and then use them to demonstrate features of the most common epigenome browsers. It should be noted that the goal is an introduction to the utility of genome browsers and therefore we are not discussing any of the available online analysis tools such as Galaxy, GREAT, DAVID, or SPARK (Giardine et al., 2005; Huang et al., 2009; Blankenberg et al., 2010; Goecks et al., 2010; McLean et al., 2010; Nielsen et al., 2012).

Epigenome Data Resources

Classically, most researchers will utilize browsers to visualize their own data sets; however, as public projects and individual investigators continue to fill the wide epigenomic space, the real and added value will be in the display and crossinvestigation of different data types that can be viewed together to generate mechanistic insights. To provide an update on major data generation projects, we will provide a brief, historically organized summary. It is advisable to be familiar with the different consortia that generate substantial amounts of epigenomic data, thus creating resources useful to researchers around the world. Figure 1 shows the current number of epigenome data sets for different data types available in each of the ongoing major projects.

Human Epigenome Project

One of the earliest organized large-scale efforts to gather epigenomic data was the Human Epigenomic Project (HEP). As part of HEP, DNA methylation data using classic bisulfite sequencing across parts of chromosomes 6, 20, and 22 spanning 873 genes were collected for 43 different tissue types between 1999 and 2006 (Rakyan et al., 2004; Eckhardt et al., 2006).

Encyclopedia of DNA Elements

The objective of the Encyclopedia of DNA Elements (ENCODE) is to catalog the functional elements in the human genome (The ENCODE Project Consortium, 2004). The data types available include gene expression, chromatin modifications, DNA methylation, chromatin accessibility, chromatin interaction, and SNP data (The ENCODE Project Consortium, 2011; Rosenbloom et al., 2012). One of the core cell types is the human ESC

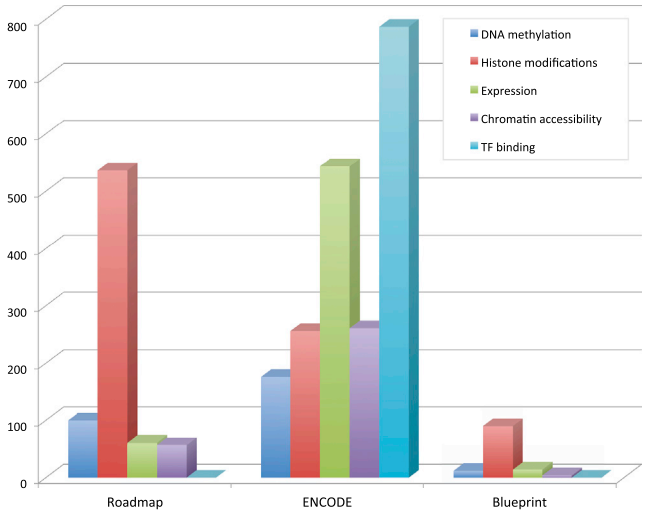

Figure 1. Epigenomic Data Sets
The number of human epigenomic data sets in each data resource (completed as of May 2013) is shown.

(hESC) line H1, also known as WA01 (http://encodeproject.org/ENCODE/).

NIH Roadmap Epigenomics Consortium

The NIH Roadmap Epigenomics Consortium aims to create a resource of epigenome reference maps in several different cell types with a particular focus on stem cells and their derivatives (Lister et al., 2009, 2011; Bernstein et al., 2010; Bock et al., 2011; Gifford et al., 2013; Xie et al., 2013). The epigenome maps consist of a few different data types: DNA methylation, six histone modifications, chromatin accessibility, and RNA expression (Figure 1; http://www.roadmapepigenomics.org).

International Human Epigenome Consortium

Building on the success of the NIH project, the International Human Epigenome Consortium (IHEC) was created to expand and coordinate worldwide epigenome mapping while simultaneously establishing universal standards for the creation of reference epigenomes. In addition to the US-based NIH Roadmap Epigenome project, it includes projects funded by the EU, Germany, Canada, and Japan (http://www.ihec-epigenomes.org).

Blueprint Epigenome

The EU-funded Blueprint Epigenome project aims to create reference epigenomes for 50 different primary human cell types and their neoplastic counterparts. The first data from the Blueprint Epigenome project were released in April 2013 and consisted of 10 complete epigenomes (http://www.blueprint-epigenome.eu).

German Epigenome Program

The German Epigenome Program (DEEP) will create 70 reference epigenomes from 2012 through 2017, specifically targeting human cell types involved in metabolic and inflammatory diseases (http://www.deutsches-epigenom-programm.de).

NCBI Epigenomics

For the past several years, scientific journals have asked that next-generation sequencing data be deposited into the NCBI Gene Expression Omnibus (GEO) repository at the time of paper publication (see http://www.nature.com/authors/policies/availability.html for an example). Therefore, the NCBI GEO contains not only all data sets released by the Roadmap Epigenomics and ENCODE projects, but also many other published epigenomic data sets (Barrett et al., 2013). In 2010, the NCBI Epigenomics Project was created to annotate these data sets and make them easier to find (http://www.ncbi.nlm.nih.gov/epigenomics). While this is an ongoing effort and data set annotation is highly dependent on the authors, the NCBI Epigenomics repository is the largest collection of epigenomics data sets currently available, as it is a superset of the Roadmap and ENCODE projects mentioned above.

(Epi)genome Browsers

Visualization and subsequent exploration of epigenome information, both from the sources described above and newly collected data sets, can be done through epigenome browsers. Some of these browsers, like Ensembl and UCSC, have existed from well before the completion of the Human Genome Project. Other browsers, such as the Human Epigenome Browser at WashU and Genboree, have been developed much more recently to specifically exploit the rich variety of genomic and epigenomic data sets being generated today, and provide specific tools to handle deeper exploration and more advanced analysis of these data sets. Table 1 contains a list of the most popular browsers that are reviewed here.

One key feature for epigenome browsers is access to the large public data repositories described above. While every browser listed in this review has the ability to import external data sets, having built-in access usually saves significant time in formatting, processing, and importing of data sets. Built-in data resource access also often implies that the browser developers have made significant efforts to ensure accuracy and usability of sample metadata and experimental variables. Given the usefulness of built-in access to data resources, we have created a matrix showing which browsers can access which data resources in a direct manner (Figure 2).

UCSC

The human genome browser at UCSC was released in order to provide access to UCSC Golden Path assemblies of the Human Genome Project along with annotation of genes and transcripts (Kent et al., 2002). It was the first browser to use the term *track* for the display of different kinds of annotation or data along the genomic sequence. Over time, the UCSC genome browser has been enhanced to include many different genomes and types of data sets.

The UCSC browser is the primary portal to data from the ENCODE project (The ENCODE Project Consortium, 2011; Rosenbloom et al., 2012). Data from the Roadmap project is also easily added by clicking the "track hubs" button in the browser.

One issue that we noted with the UCSC Genome Browser is that track colors for built-in data sets are not configurable but are hard-coded by the data publishers, such as ENCODE. While this is rarely an issue for tracks originating for the same data resource (as the publisher assigns consistent track coloring), it can be an issue when comparing data sets from two different data resources, where one might want to color epigenetic data consistently by type (see for instance the DNA methylation data in Figure 3). One workaround is to download the track data from the UCSC browser, reupload it as a custom track,

Table 1. List of Epigenome Browsers

Name	URL	Platform	Unique Features
Ensembl	http://www.ensembl.org	web	
UCSC	http://www.epigenomebrowser.org/	web	
Sequence viewer	http://www.ncbi.nlm.nih.gov/epigenomics	web	basic statistical analyses
Human epigenome browser at WashU	http://epigenomegateway.wustl.edu/	web	gene set view, gene plots
IGV	http://www.broadinstitute.org/igv/	desktop	interface to GenomeSpace

and then change the color. If more customization is required, we recommend using the Integrative Genome Viewer (IGV) or the WashU Epigenome browser, which have more sophisticated track customization options.

Ensembl

The first attempt to show annotation along a genomic sequence was the Ensembl genome database (Hubbard et al., 2002). Since Ensembl was designed to handle draft genome sequences with assembly being a work in progress, the fundamental unit of the Ensembl genome browser was the *virtual contig*. Gene structures were initially added using the Genscan algorithm. Other annotations such as ESTs and SNPs were also displayed. Ensembl served the important function of being a reference database during early genome sequencing efforts.

Ensembl has been continuously updated since its initial release, and the latest versions include several other kinds of annotation and data. Extensive documentation on the use of the Ensembl genome browser is available both in previous publications (Spudich and Fernández-Suárez, 2010) and on the Ensembl website. Epigenomic data available in the Ensembl genome browser include released data from the Roadmap and ENCODE projects.

NCBI Sequence Viewer

The GEO is NCBI's analog to GenBank for expression and high-throughput sequencing data sets (Barrett et al., 2013). Epigenomic data sets submitted to GEO can be viewed in a genomic context using the Sequence Viewer browser (in addition to the UCSC Genome Browser) through the NCBI Epigenomics Portal (http://www.ncbi.nlm.nih.gov/epigenomics).

Human Epigenome Browser at WashU

The Human Epigenome Browser at WashU is intended to be a browser specifically for viewing epigenomic and next-generation sequencing data sets (Zhou et al., 2011). It provides built-in access to the Roadmap Epigenome and ENCODE data sets, with the ability to define other "Data Hubs."

One distinctive feature of the WashU browser is the Gene Set View. The Gene Set View shows several discrete regions of the genome simultaneously. These regions need not be defined in terms of genomic coordinates, but can be chosen to center on the transcription start sites (TSSs) of a set of genes. The set of genes can also be chosen using KEGG pathway annotation.

Another unique feature of the WashU browser is the Gene Plot. Given a data track and a set of genomic regions, the browser divides the genomic regions into a specified number of windows and draws a boxplot of the values for the track over these windows. The Gene Plot feature allows the user to visually see the distribution of track values across the length of genomic regions and can also be used to compare these distributions across tracks (see Figure 5 for an example).

Integrative Genome Viewer

Unlike the other browsers in this review, the IGV is a desktop application written in Java (Robinson et al., 2011). While the number of built-in data sources is limited to an infrequently updated snapshot of ENCODE, IGV can load data from a local file, a file located online, or a Distributed Annotation System (DAS) server such as UCSC. The advantages of IGV are its responsiveness, its performance on large data sets, and its ability to extensively customize how tracks are viewed. At the same time, performance with remote data sets is much less smooth, making IGV most useful for viewing locally generated epigenomic data.

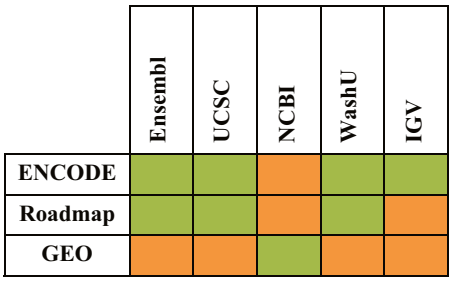

Figure 2. Matrix of Data Sources and Browsers Showing the Availability of Each Epigenome Data Source on Each Epigenome Browser

Green cells indicate that there is an instance of the browser already set up with the data source included. Orange indicates that data files from the data source must be imported manually into the browser. Note: GEO also contains data from ENCODE and the Roadmap Epigenomics project.

The Utility of Epigenome Browsers

Similar to the wide use of the genome browsers for exploring sequence conservation, gene location, and other genomic information, the epigenome can be used to quickly explore and gather information that would have taken a long time in the pre-epigenome era. In the following sections, we will illustrate this utility with several stem-cell-centric examples.

Quality Control and Data Processing

As with any experiment, the creation of epigenomic data sets is often done in replicates, either technical or biological. Comparing data sets generated from replicates can yield a useful metric of data quality. We used the Table Browser function of the UCSC Genome Browser to compare DNA methylation data in H1 ESCs measured by reduced representation bisulfite sequencing across two replicates from the ENCODE project. The overall correlation coefficient was 0.77, and the output includes the correlation coefficients for each chromosome (Figures S1 and S2).

A user might also want to compare epigenomic data sets created using two different platforms. To demonstrate this we

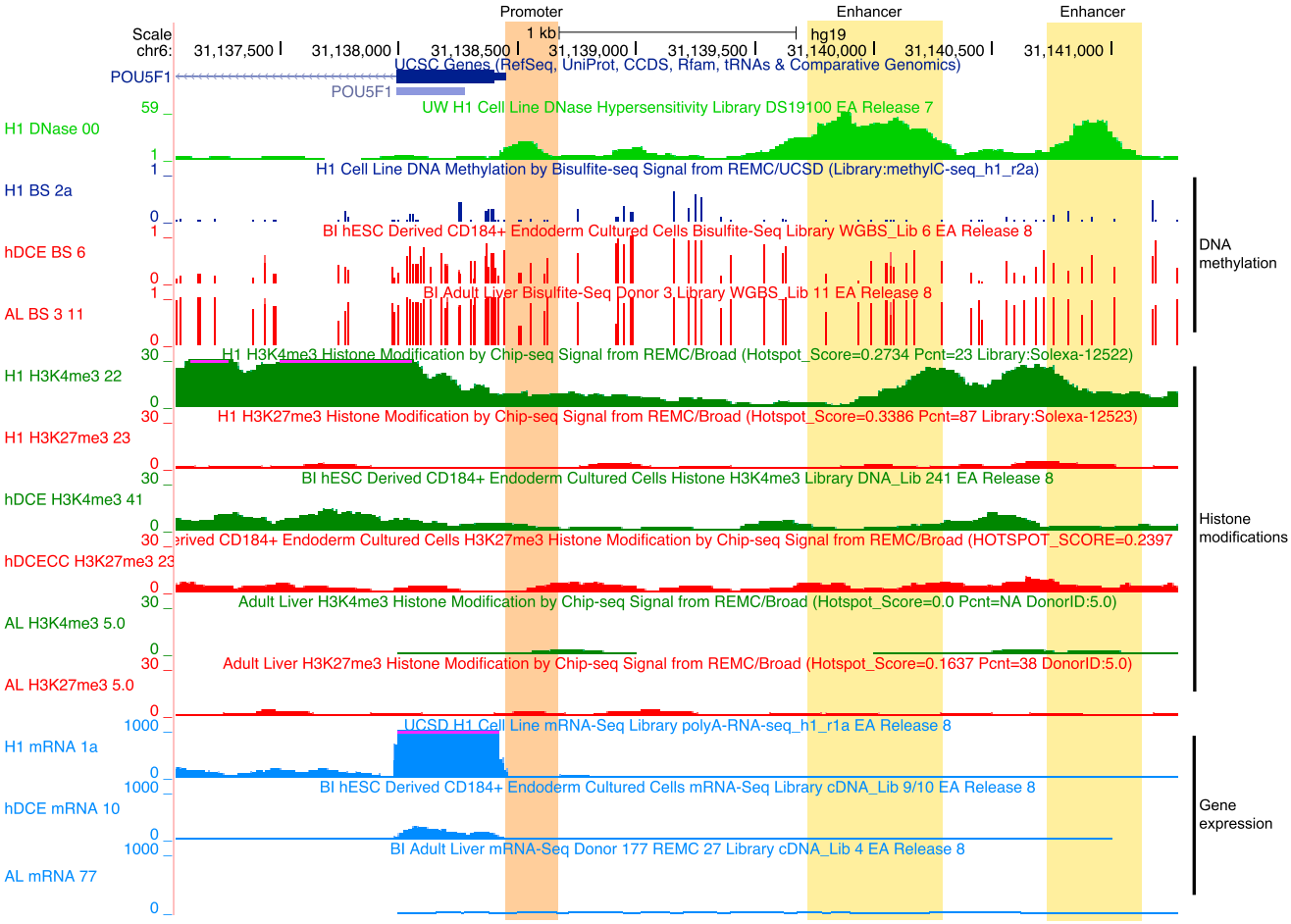

Figure 3. The Epigenomic and Regulatory Landscape of *OCT4/POU5F1* in ESCs, Endoderm Cells, and Adult Liver

The promoter, proximal enhancer, and distal enhancer are highlighted using the DNase hypersensitivity track (light green) as a guide. All three regulatory elements are hypomethylated in ESCs and hypermethylated in the differentiated cells (red barplots). H3K4me3 levels correlate with the expression of OCT4. DNA methylation rather than H3K27me3 appears to confer long-term silencing of the locus. Data are taken from Roadmap Epigenomics project and viewed in UCSC Genome Browser. To utilize browser images like this one for presentation/publication, additional visual improvements (labels, adjusted font sizes, colors, etc.) are recommended.

compared DNA methylation data in H1 ESCs measured by whole-genome bisulfite sequencing (WGBS) as part of the Roadmap Epigenomics project and the Human Infinium Methylation 450 bead array as part of ENCODE over approximately 150 kb of chromosome 6. The overall correlation between the data sets is very good, and both platforms identify the same hypomethylated regions (Figure S3).

Identify Locus-Level Mechanisms of Regulation

One of the most common uses of epigenome browsers is in viewing epigenetic information that might regulate the expression of a single gene. Figure 3 shows such a view of the pluripotency gene *OCT4/POU5F1* in several different cell types as viewed through the UCSC browser. The data sets here are from the Roadmap Epigenomics project. In order to access tracks from this project in the UCSC browser, use the "track hubs" button and enable the Roadmap track hub. Once this track hub is enabled, the Roadmap tracks can be added to the browser using the dropdowns in the "Roadmap Epigenomics Data Complete Collection at Wash U VizHub" section of the page. You can choose data tracks for specific samples and assays by clicking on the links above the dropdowns. Other top-level sections on the page allow access to ENCODE consortium data.

One can immediately make several interesting observations about the possible regulation of *OCT4* simply using the browser view in Figure 3. This view shows DNase hypersensitivity, DNA methylation, histone modifications, and gene expression in the upstream region of *OCT4* in three different human cell types: H1 ESCs, CD184+ endoderm cells derived from HUES64 ESCs, and adult liver cells. First, we can see that the DNase hypersensitivity track has clear peaks at the proximal promoter and the two enhancer elements that are known to regulate the expression of *OCT4*. Second, the promoters and enhancers are hypomethylated in the pluripotent H1 cells, but hypermethylated in the endoderm and liver cells. Third, we also see high levels of H3K4me3 (which marks active/poised genes) and low levels of H3K27me3 (a repressive modification) in this region in H1 cells, but low levels of H3K4me3 in the endoderm and liver cells. It is known that the *OCT4* upstream regions acquire first H3K9me2 and then DNA methylation after transcriptional silencing has already occurred (Epsztejn-Litman et al., 2008; Athanasiadou

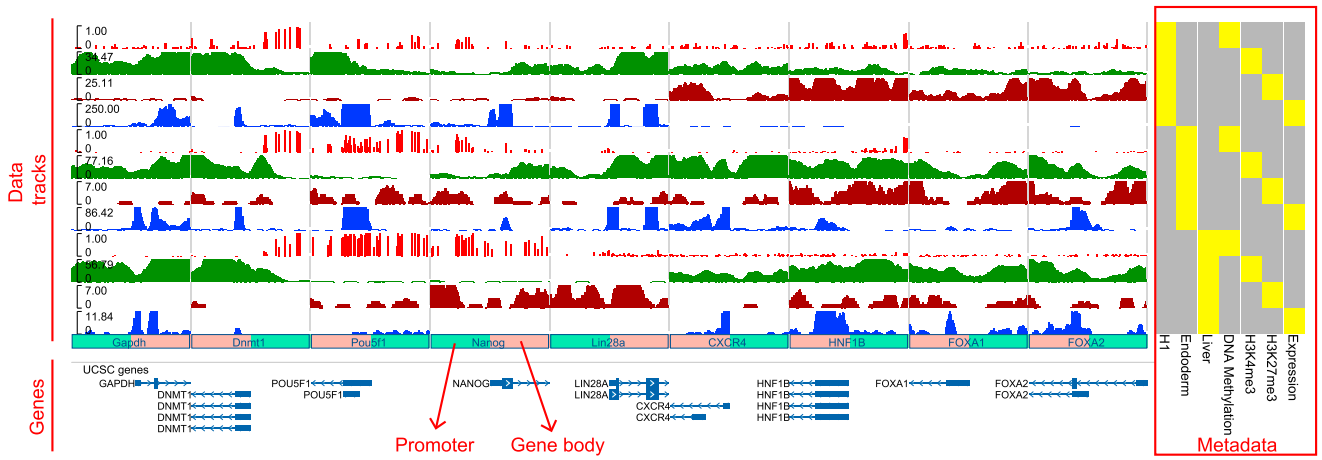

Figure 4. A Gene Set View in the WashU Epigenome Browser

The columns on the right side represent track annotation: a yellow highlight means that the annotation term applies to the track, e.g., the first four tracks show data in H1 ESCs; the first, fifth, and ninth tracks are DNA methylation; and so on. The genomic regions shown are 1,000 bp on either side of the transcription start site (TSS) of the set of genes, which belong to four groups: housekeeping genes (*GAPDH* and *DNMT1*), pluripotency genes regulated by DNA methylation (*OCT4*/*POU5F1*, *NANOG*) or H3K27me3 (*LIN28A*), and bivalent genes (*CXCR4*, *HNF1B*, *FOXA1*, and *FOXA2*). For each gene, the upstream region/promoter is marked in green and the gene body (region downstream of TSS) is marked in red. Data are from Roadmap Epigenomics project, viewed in the Epigenome Browser at WashU.

et al., 2010). The high levels of DNA methylation can be observed in the adult liver, but the endoderm population shows somewhat lower intermediate levels for most CpGs and, interestingly, slightly increased levels of H3K27me3. These two marks strongly influence each other's presence and are mutually exclusive in CpG-rich contexts (Brinkman et al., 2012). Whether the H3K27me3 signal represents a true transient signal or is for example increased background noise cannot be assessed from this simple visual analysis and would require additional experimental validation. Nonetheless, it is an interesting example of hypothesis generation from simple genome browser views. Finally, the bottom tracks show the expected trend of high *OCT4* expression in the stem cells and reduced expression in the endoderm and no detectable expression in liver cells (Figure 3).

While much of the above information has been collected from studies over the past decade, this display nicely demonstrates how easy it is these days to obtain mechanistic insights into the regulation of specific genes by overlaying information from several different assays in epigenome browsers. This information is generally also useful when designing locus-specific primers for measuring DNA methylation or ChIP-qPCR or cloning the upstream regulatory region of a given gene as the exact positions of putative regulatory elements are clearly delineated. The design of such experimental tools was done "blindly" on the genome sequence in the past, but now one can specifically target them to the most dynamic and hence likely relevant regions.

Our second example is taken from the mouse cardiac lineage, which was profiled recently for histone modifications and gene expression (Wamstad et al., 2012). As mentioned in that study, each stage in the lineage exhibits expression of specific markers: *Nanog* for pluripotent stem cells, *Mesp1* for cells differentiated into the mesoderm lineage, *Nkx2-5* for cardiac precursor cells, and *Myh6* for functional cardiomyocytes. We downloaded the BigWig files published by the authors and used IGV to display the histone modifications and expression around these four genes across four different cell types (Figure S3). The expression patterns of the genes map to the cell types as expected, and we can see that H3K27ac and H3K4me3 are generally correlated with expression. *Nkx2-5* is not expressed in stem cells, gains bivalent H3K4me3 and H3K27me3 marking (i.e., is "poised to activate" as noted by the authors) in mesoderm cells, and subsequently loses H3K27me3 and is expressed in cardiomyocytes.

A third example shows the expression of the muscle gene *MYOD1* across several different cell lines measured as part of the ENCODE project (Figure S4). As we would expect, *MYOD1* is only expressed in human skeletal muscle myoblasts (HSMMs) and not in Gm12878, H1, HUVEC, NHEK, and NHLF. In keeping with this observation, H3K4me3, generally associated with active genes, is enriched only in HSMM cells at this locus.

Identify Trends in Sets of Genes

Though looking at epigenetic marks at a single genomic locus can be instructive, a mechanistic trend often becomes more obvious when visualizing a set of loci. The Human Epigenome Browser at WashU has two views, the Gene Set View and the Gene Plot, that can show trends across several genomic loci.

The Gene Set View lets the user define a set of genes or genomic regions of interest and view epigenetic information about these loci side by side. The browser allows the user to either define a set of genomic regions centered on annotated genes (for example, 1,000 bp on either side of the TSS) or enter a set of genomic coordinates directly. The Gene Set View can be accessed using the Apps tab of the browser's floating menu.

We show an example of a Gene Set View in Figure 4. There are four different groups of genes shown side by side: housekeeping genes (*GAPDH* and *DNMT1*), pluripotency genes regulated by DNA methylation (*OCT4*/*POU5F1* and *NANOG*), a pluripotency gene regulated by H3K27me3 (*LIN28A*), and bivalent genes (*CXCR4*, *HNF1B*, *FOXA1*, and *FOXA2*). The epigenetic information shown is DNA methylation, H3K4me3, H3K27me3, and gene expression in three different human cell types, H1 ESCs, HUES64 derived endoderm cells, and adult liver cells. We can see that the housekeeping genes *GAPDH* and *DNMT1* are

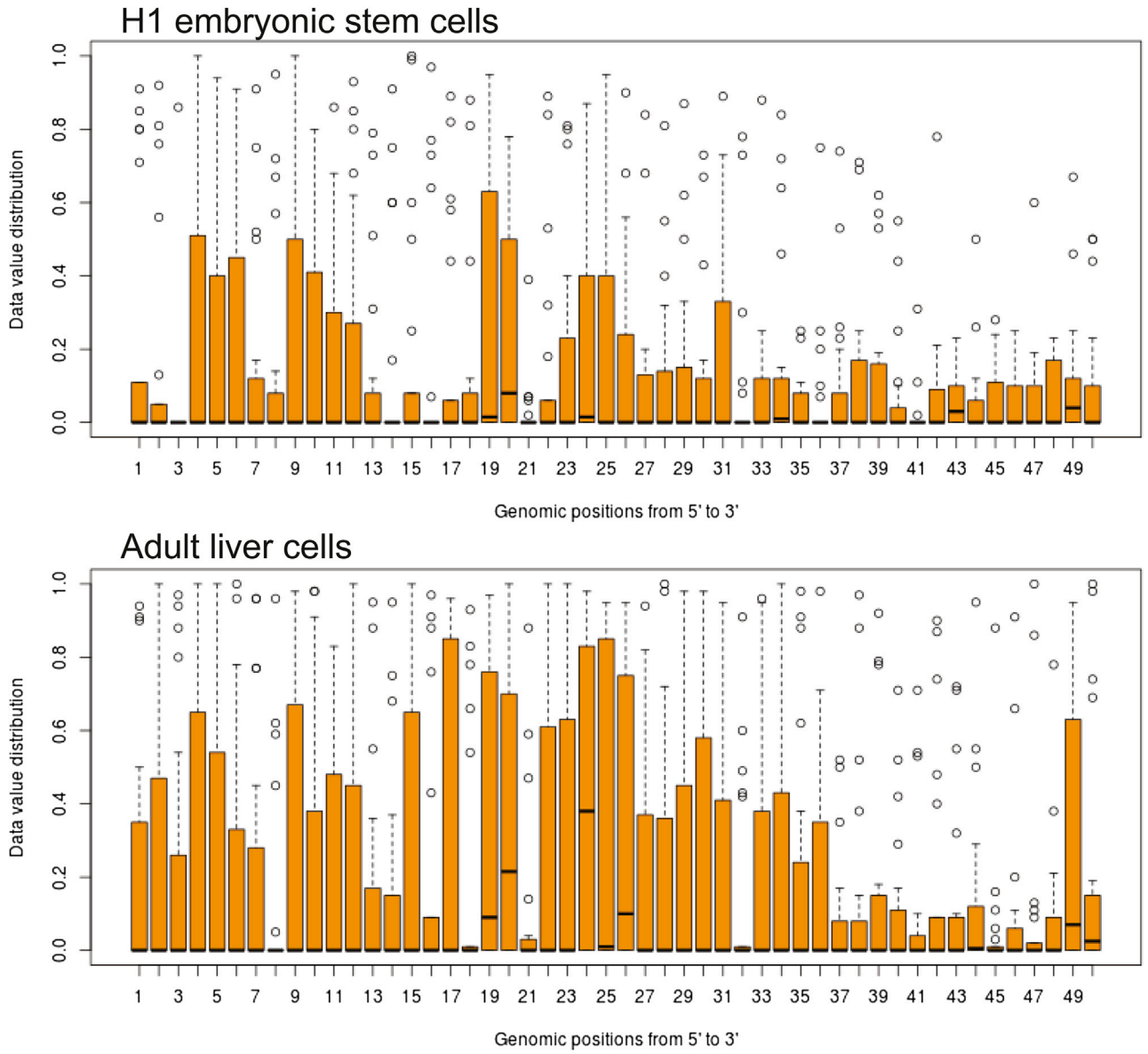

Figure 5. Comparison of DNA Methylation Levels over 1 kb of the Promoter of 22 Genes in H1 ESCs (Top) and Adult Liver (Bottom)
These promoters were identified as hypomethylated in iPSCs and hypermethylated in differentiated cells by Nishino et al. (2010). Data are from Roadmap Epigenomics project and plots were created using WashU Epigenome Browser. The box plot has hinges at the first and third quartiles, and the whiskers mark 1.5 times the interquartile range.

expressed in all three cell types: they show low DNA methylation, high H3K4me3, and low H3K27me3 at the TSS. The pluripotency genes *OCT4/POU5F1* and *NANOG* are expressed in H1 cells, but they are silenced and marked by DNA methylation in differentiated cells. *LIN28A* is repressed and marked by H3K27me3 in differentiated cells but is expressed in H1 cells where this mark is absent. Finally, the bivalent genes have overlapping patterns of H3K4me3 and H3K27me3 in H1 cells, but lose H3K27me3 and are expressed in the differentiated cells.

Instead of a gene-centric approach, the Gene Set View can also be used to view up to 200 genomic regions. For example, we looked at DNA methylation at the 100 largest experimentally verified CpG islands (Illingworth et al., 2010) on the X chromosome in ESCs and iPSCs sorted by gender (Figure S5). This visualization clearly shows that there is substantial variation in the DNA methylation levels, in particular among female lines as previously reported (Mekhoubad et al., 2012).

We used the Gene Plot feature of the WashU browser to compare the DNA methylation levels at the promoters of 22 genes among ESCs and adult liver. The Gene Plot divides the region of interest into a user-specified number of windows and creates boxplots of any quantitative track on a window-by-window basis. The genes we chose were previously identified as having hypomethylated promoters in iPSCs and hypermethylated promoters in the parent somatic cells (Nishino et al., 2010), and the gene plots clearly show this trend (Figure 5).

A Caveat about Data Processing and Normalization

As seen from the examples above, epigenome browsers allow a user to quickly view data collected by many different labs and/or projects side by side. While comparing such data, it is important to keep in mind that the publishers might have used different data processing and normalization procedures. In most of the browsers, qualitative descriptions of these processes can be viewed quite easily to get a basic understanding of what methods were used. In addition, each epigenome-mapping project has designated standards for these activities. We recommend that epigenome browsers be used to identify trends and generate hypotheses but also that a statistically rigorous comparison including reprocessing and renormalization of the raw data be done before definitive conclusions are drawn.

An Epigenomic Workflow: Visualize, Hypothesize, and Validate

We have shown above how epigenome browsers can be utilized in two separate parts of a typical epigenomic workflow. First, after generation of raw epigenomic data sets, epigenome browsers allow bench scientists to quickly determine the quality of these data sets before any downstream analysis. Second, we can visualize processed data sets in epigenome browsers to identify interesting local trends and generate predictions (hypothesize) about regulatory processes operating at a genome-wide level. We can then validate these predictions using more advanced bioinformatic analyses. While statistical validation of hypotheses is the important final step in the epigenomic workflow and has yielded vital insights about genomic and epigenomic regulation (Bernstein et al., 2006; Chen et al., 2008; Guttman et al., 2011), the specific methods used can vary widely depending on the hypotheses being tested and fall outside the scope of this review.

Outlook and Conclusion

A major goal for the massive and coordinated epigenome efforts is to advance our understanding of how the epigenome contributes to the specification and maintenance of diverse cell types. This improved understanding of the functional significance of the epigenome will in turn provide insights into disease biology and thereby eventually impact human health, in line with the ultimate objectives of the Human Genome Project and its natural NIH funded offshoots, the ENCODE project and the Roadmap Epigenomics Consortium. A growing number of genomes and epigenomes are used in a daily fashion that may not be obvious from publications, but are nonetheless critical contributors to biological research today. Researchers will soon use epigenome data to both check whether their gene of interest is expressed in particular cell types and how it is regulated and precisely define where its functional and regulatory elements are located. Genome browsers that can facilitate accessing and visualization of epigenomic data sets are crucial to maximizing the scientific value of these important data resources.

SUPPLEMENTAL INFORMATION

Supplemental Information for this article includes five supplemental figures and can be found with this article online at http://dx.doi.org/10.1016/j.stem.2013.06.006.

ACKNOWLEDGMENTS

We thank Xin Zhou, Ting Wang, and Ian Fingerman for helpful suggestions on the manuscript. A.M. is supported by the Pew Charitable Trusts and US National Institutes of Health (NIH) grants (U01ES017155 and P01GM099117) and is a Robertson Investigator of The New York Stem Cell Foundation.

REFERENCES

Athanasiadou, R., de Sousa, D., Myant, K., Merusi, C., Stancheva, I., and Bird, A. (2010). Targeting of de novo DNA methylation throughout the Oct-4 gene regulatory region in differentiating embryonic stem cells. PLoS ONE *5*, e9937.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*(Database issue), D991–D995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel … [et Al.] *Chapter 19*, Unit 19.10.1–21.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell *144*, 439–452.

Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., and Stunnenberg, H.G. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. *22*, 1128–1138.

Chadwick, L.H. (2012). The NIH Roadmap Epigenomics Program data resource. Epigenomics *4*, 317–324.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet. *38*, 1378–1385.

ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H., and Bergman, Y. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat. Struct. Mol. Biol. *15*, 1176–1183.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., et al. (2013). Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. Cell *153*, 1149–1163.

Goecks, J., Nekrutenko, A., and Taylor, J.; Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the Mamm. Genome. PLoS Genet. *6*, 15.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature *471*, 68–73.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. Nat. Biotechnol. *28*, 1079–1088.

Mekhoubad, S., Bock, C., de Boer, A.S., Kiskinis, E., Meissner, A., and Eggan, K. (2012). Erosion of dosage compensation impacts human iPSC disease modeling. Cell Stem Cell *10*, 595–609.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Nielsen, C.B., Younesy, H., O'Geen, H., Xu, X., Jackson, A.R., Milosavljevic, A., Wang, T., Costello, J.F., Hirst, M., Farnham, P.J., and Jones, S.J. (2012). Spark: a navigational paradigm for genomic data exploration. Genome Res. *22*, 2262–2269.

Nishino, K., Toyoda, M., Yamazaki-Inoue, M., Makino, H., Fukawatase, Y., Chikazawa, E., Takahashi, Y., Miyagawa, Y., Okita, H., Kiyokawa, N., et al. (2010). Defining hypo-methylated regions of stem cell-specific promoters in human iPS cells derived from extra-embryonic amnions and lung fibroblasts. PLoS ONE *5*, e13017.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G.A., Stewart, R., and Thomson, J.A. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell *1*, 299–312.

Rakyan, V.K., Hildmann, T., Novik, K.L., Lewin, J., Tost, J., Cox, A.V., Andrews, T.D., Howe, K.L., Otto, T., Olek, A., et al. (2004). DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol. *2*, e405.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H., et al. (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res. *40*(Database issue), D912–D917.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. *14*, 204–220.

Spudich, G.M., and Fernández-Suárez, X.M. (2010). Touring Ensembl: a practical guide to genome browsing. BMC Genomics *11*, 295.

The ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements). Project. Science *306*, 636–640.

Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell *151*, 206–220.

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., et al. (2013). Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell *153*, 1134–1148.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. Nat. Methods *8*, 989–990.